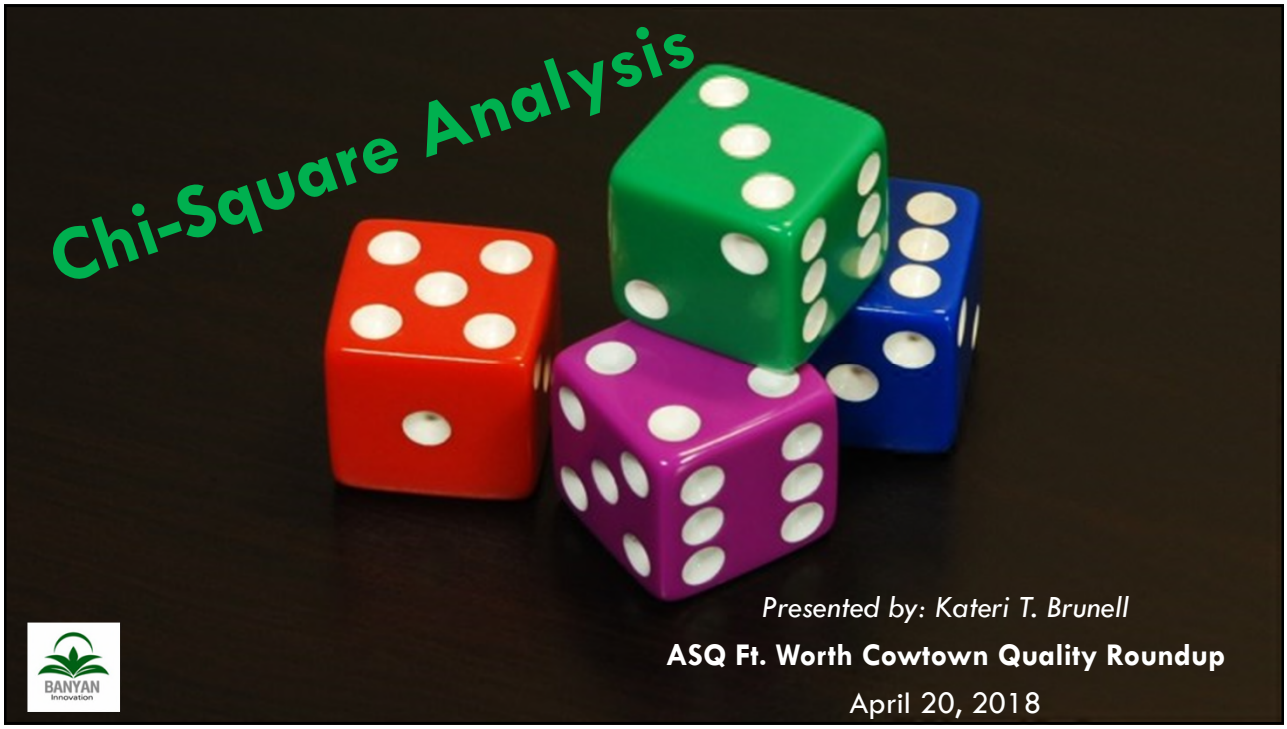
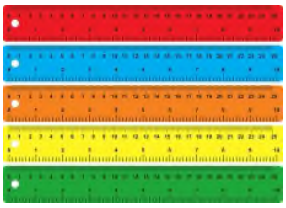


Learning Objectives

- Review concepts of categorical data and contingency tables
- Discuss Chi Square distribution
- Review example for 2x2 Chi Square test
- Discuss Goodness-of-Fit test (>2x2 tables)
- Explain how to interpret the results of a Chi Square test
- Demonstrate how to use Excel, SigmaXL software

Two Different Types of Data



Attribute (Counted)

- Counted
- Proportion of occurrence (p)
- Use Z-distribution
- *Examples: defectives*

Variables (Measured)

- Measured
- Variance or Standard deviation (s)
- Uses t- or Z distribution
- *Example: Average cost of defects*

Attribute Data

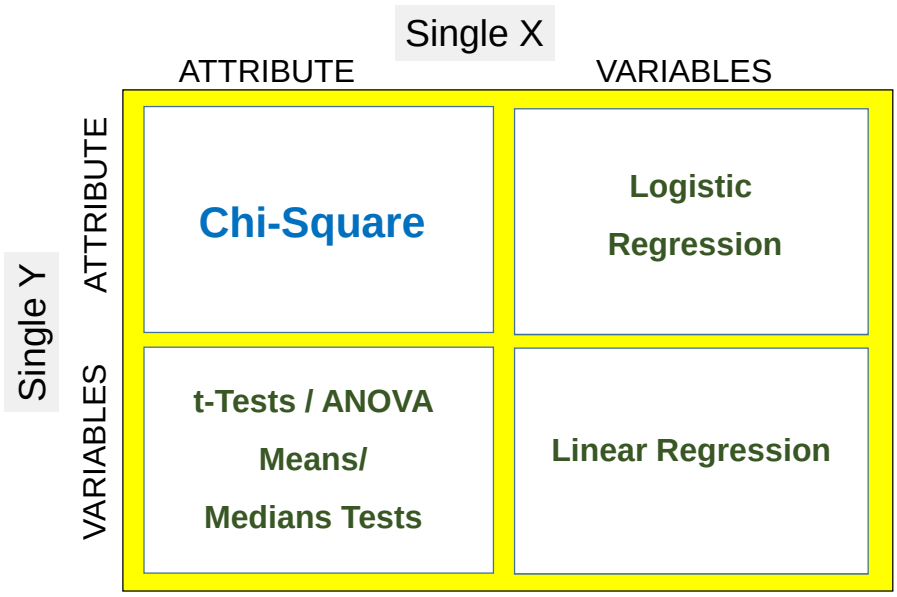
- Dichotomous

- Characteristics

- Nominal Categories

- Ordinal Categories

When to Use Chi-Square



Chi-Square Example (2 x 2 Table)

A company piloted a new product with 10,000 customers. After the pilot, they wanted to determine the customers' likelihood of buying the product again within the next month ("Yes or No" checked on a brief survey).

They got results from an initial sample of 480 customers: 195 new and 285 existing customers

Now they want to determine if there is a significant relationship between the customer type and their likelihood of repurchase.

(X=Customer Type: Old/New, Y=Repurchase Y/N)

Customer Type	Repurchase
Old	Yes
New	No
New	No
Old	Yes
New	Yes
Old	No
Old	No
Old	Yes
New	Yes
Old	No
Old	No
New	No
New	No
Old	Yes
Old	No
Old	No
Old	No
Old	Yes
Old	Yes
Old	Yes
Old	No
New	No

Contingency Table

First, build a contingency table:

- easily done using pivot table in Excel
- aka Crosstabulation or Crosstab

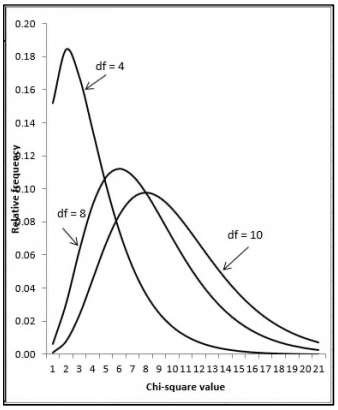
Repurchase? / Customer Type	No	Yes	Grand Total
New	136	59	195
Old	156	129	285
Grand Total	**292**	**188**	**480**

- This gives us our observed values

Can we draw any conclusions about each customer type's intent to repurchase?

Need to conduct a statistical test for independence

Chi-Square Distribution



- Shape depends on degrees of freedom (df)
- Df determined by number of rows and columns
- Df = (r-1) * (c-1), where:
  - R is number of rows
  - C is number of columns
- Compares the **observed frequencies** to the **expected frequencies** (what we would expect to see in each cell if the two data sets were independent)

Mean = df
Variance = 2df

$$\chi^2 = \frac{\sum(f_o - f_e)^2}{f_e}$$

where:
f_o = Observed frequency
f_e = Expected frequency

Chi-Square Test – Step 1

- State hypotheses:

H_o: Data are Independent (Not Related) → Repurchase Intent is Not Related to Customer Type

H_a: Data are Dependent (Related) → Repurchase Intent is Related to Customer Type

- Set alpha level (decision cutoff point)

  If the P Value is <.05 , then reject H_o

Chi-Square Test – Step 2

- Compile data for observed frequencies, including row and column totals

Repurchase? / Customer Type	No	Yes	Grand Total
New	136	59	195
Old	156	129	285
Grand Total	**292**	**188**	**480**

This is the same as the contingency table that we created earlier

Chi-Square Test – Step 3

- Calculate expected frequencies for each cell (second table)

Repurchase? / Customer Type	No	Yes	Grand Total
New	136	59	195
Old	156	129	285
Grand Total	**292**	**188**	**480**

Cell's expected frequency is:

(Row Total) X (Column Total)
Grand Total

	NO	YES	GRAND TOTAL
NEW	**195*292 / 480 =118.63**	195*188 / 480 = 76.37	195
OLD	173.37	111.63	285
GRAND TOTAL	292	188	480

Chi-Square Test – Step 4

- Subtract expected from the observed frequencies for each cell (third table)

Repurchase? / Customer Type	No	Yes	Grand Total
New	136	59	195
Old	156	129	285
Grand Total	**292**	**188**	**480**

New cell values are:
(O - E)

	NO	YES
NEW	**136-118.63 = 17.37**	59-76.37 = (17.37)
OLD	156-173.37 = (17.37)	129-111.63 = 17.37

Chi-Square Test – Step 5

- Square the differences (fourth table)

	NO	YES
NEW	**136-118.63 = 17.37**	59-76.37 = (17.37)
OLD	156-173.37 = (17.37)	129-111.63 = 17.37

New cell values are:
$(O - E)^2$

	NO	YES
NEW	**17.37*17.37 = 301.72**	(17.37)* (17.37) = 301.72
OLD	(17.37)* (17.37) = 301.72	17.37*17.37 = 301.72

Chi-Square Test – Step 6

- Compute the relative squared differences (fifth table)

	NO	YES
NEW	**17.37*17.37 = 301.72**	(17.37)* (17.37) = 301.72
OLD	(17.37)* (17.37) = 301.72	17.37*17.37 = 301.72

New cell values are:
$\frac{(O - E)^2}{E}$

	NO	YES
NEW	**301.72 / 118.63 = 2.54**	301.72/ 76.37 = 3.95
OLD	301.72 / 173.37 = 1.74	301.72 / 111.63 = 2.70

Chi-Square Test – Step 7

- The sum of the relative squared differences is the Chi Square distribution

	NO	YES
NEW	2.54	3.95
OLD	1.74	2.70

2.54 + 3.95 + 1.74 + 2.70 = **10.93**

- If there is independence, we expect total Chi-Square to be close to 0
- The larger the number, the more likely the variables are dependent
- However, we must use the **P value** to make the final determination

Chi-Square Test – Step 8

- Remember our hypotheses from Step 1:

H_o: Data are Independent (Not Related) → Repurchase Intent is Not Related to Customer Type

H_a: Data are Dependent (Related) → Repurchase Intent is Related to Customer Type

- Set alpha level (decision cutoff point)

If the **P Value is <.05 , then reject H_o**

Chi-Square Test – Step 8 (cont'd)

- Determine degrees of freedom (r-1)* (c-1) = (2-1)* (2-1) = **1**
- Use Excel's Chi-Square test function to determine p-value

	NO	YES
NEW	2.54	3.95
OLD	1.74	2.70
Chi Square		10.93
P-Value		0.00094619

=CHISQ.DIST.RT(Cell Reference for Total Chi-Square, Degrees of Freedom)

=**CHISQ.DIST.RT**(**10.93, 1**)

If the P Value is <.05 , then reject H_o

Conclusion: Repurchase intent is dependent on customer type

Further Analysis

- The Chi-Square test only determines if there is a relationship
- Does not show cause and effect
- Need to examine data in the table itself (conditional probabilities)

Chi-Square Test
Customer Type - Repurchase

Observed	No	Yes	
New	136	59	195
Old	156	129	285
	292	188	

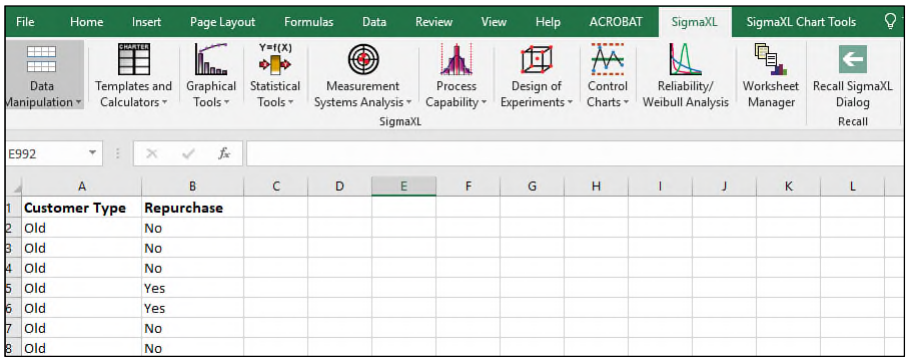
Expected	No	Yes	
New	118.63	76.37	195
Old	173.37	111.63	285
	292.00	188.00	

Chi-Square Test
Customer Type - Repurchase

Observed	No	Yes	
New	0.70	0.30	1
Old	0.55	0.45	1
	1.24	0.76	

Expected	No	Yes	
New	0.61	0.39	1
Old	0.61	0.39	1
	1	1	

Example Using Excel (SigmaXL)



	A	B
1	**Customer Type**	**Repurchase**
2	Old	No
3	Old	No
4	Old	No
5	Old	Yes
6	Old	Yes
7	Old	No
8	Old	No

Jump to Excel

Goodness-of-Fit Test

- Tests if sample data fits a specific population distribution (model)
- Can also test the homogeneity of proportions
- Compares observed to expected frequencies in a sample
- Degrees of freedom (df) = k-1 (where k is the number of levels in the category)
- Same Chi-Square formula as with 2x2 tables

$$\chi^2 = \frac{\sum(f_o - f_e)^2}{f_e}$$

where:
f_o = Observed frequency
f_e = Expected frequency

Goodness-of-Fit Test - Example

- A plastics decorator (printer) runs three lines (A, B, and C) for imprinting artwork on cups. A total of 285 defective cups is tallied for a week.
  - Line A - 75
  - Line B - 114
  - Line C - 96
- Is the proportion defective equal across all 3 lines? (Assume the same amount of total output per line).

- What is the total Chi-Square?
- What are the degrees of freedom?

Goodness-of-Fit Test – Example (cont'd)

H_o: Defectives are equally proportional across all 3 lines
H_a: Defectives are not equally proportional across all 3 lines

	Line A	Line B	Line C	*Total*
Observed Frequencies	75	114	96	285
Expected	95	95	95	
Differences	(20)	19	1	
Differences^2	400	361	1	
Diff^2 / Expected	4.21	3.80	0.01	
Chi-Square	8.02			
Degrees of Freedom	2			
P- Value	0.018123854			

Line B has a significantly higher number of defectives

Other Chi-Square Considerations

- Larger sample sizes (>500) can be sensitive to small differences
- Expected frequencies for each cell must be at least 1, preferably >5
- Beware of hidden "lurking" third variable
- Variables data will provide more information with fewer samples….
- But Chi Square can be an effective tool with attribute data

Summary

✓Reviewed concepts of categorical data and contingency tables
✓Discussed Chi Square distribution
✓Reviewed example for 2x2 Chi Square test
✓Discussed Goodness-of-Fit test (>2x2 tables)
✓Explained how to interpret the results of a Chi Square test
✓Demonstrated how to use Excel, SigmaXL software

Questions?

For more information and additional resources:

www.banyan-innovation.com
kateri@banyan-innovation.com

(561) 352-5070